

Inference-Time Intervention: Eliciting Truthful Answers from a Language Model

Year: 2023 | Citations: 790 | Authors: Kenneth Li, Oam Patel, Fernanda Vi'egas

Abstract

We introduce Inference-Time Intervention (ITI), a technique designed to enhance the "truthfulness" of large language models (LLMs). ITI operates by shifting model activations during inference, following a set of directions across a limited number of attention heads. This intervention significantly improves the performance of LLaMA models on the TruthfulQA benchmark. On an instruction-finetuned LLaMA called Alpaca, ITI improves its truthfulness from 32.5% to 65.1%. We identify a tradeoff between truthfulness and helpfulness and demonstrate how to balance it by tuning the intervention strength. ITI is minimally invasive and computationally inexpensive. Moreover, the technique is data efficient: while approaches like RLHF require extensive annotations, ITI locates truthful directions using only few hundred examples. Our findings suggest that LLMs may have an internal representation of the likelihood of something being true, even as they produce falsehoods on the surface.